

# Large Alphabet Compression and Predictive Distributions through Poissonization and Tilting

Xiao Yang, *Student member, IEEE* Andrew R. Barron, *Fellow, IEEE*

## Abstract

This paper introduces a convenient strategy for coding and predicting sequences of independent, identically distributed random variables generated from a large alphabet of size  $m$ . In particular, the size of the sample is allowed to be variable. The employment of a Poisson model and tilting method simplifies the implementation and analysis through independence. The resulting strategy is optimal within the class of distributions satisfying a moment condition, and is close to optimal for the class of all i.i.d distributions on strings of a given length. Moreover, the method can be used to code and predict strings with a condition on the tail of the ordered counts. It can also be applied to distributions in an envelope class.

## Index Terms

large alphabet, minimax regret, normalized maximum likelihood, Poisson distribution, power law, universal coding, Zipf's law

## I. INTRODUCTION

**L**ARGE alphabet compression and prediction problems concern understanding the probabilistic scheme of a huge number of possible outcomes. In many cases the ordered probability of individual outcomes displays a quickly falling shape, with a small number of outcomes happening most often. An example is Chinese characters. A recent published dictionary contains 85568 Chinese characters in total [1], but the number of frequent characters is considerably smaller. Here we consider an i.i.d model for this problem. Despite the possible dependence among the symbols in the alphabet as in language, it serves as a start and can be extended to models taking dependence into account.

Previous theoretical analysis usually assumes the length of a message is known in advance when it is coded. This is not always true in practice. Serialization writers do not know how many words a novel contains exactly before he finishes the last sentence. Nevertheless, given a limited time or space, one could possibly guess how many words on average can be accommodated.

Suppose a string of random variables  $\underline{X} = (X_1, \dots, X_N)$  is generated independently from a discrete alphabet  $\mathcal{A}$  of size  $m$ . We allow the string length  $N$  to be variable. A special case is when  $N$  is given as a fixed number, or it can be random. In either case,  $\underline{X}$  is a member of the set  $\mathcal{X}^*$  of all finite length strings

$$\begin{aligned}\mathcal{X}^* &= \bigcup_{n=0}^{\infty} \mathcal{X}^n \\ &= \bigcup_{n=0}^{\infty} \{x^n = (x_1, \dots, x_n) : x_i \in \mathcal{A}, i = 1, \dots, n\}.\end{aligned}$$

Our goal is to code/predict the string  $\underline{X}$ . Note that the length  $N$  is determined by the string. There will be an agreed upon distribution of  $N$ , perhaps Poisson or deterministic.

Now suppose given  $N$ , each random variable  $X_i$  is generated independently according to a probability mass function in a parametric family  $\mathcal{P}_\Theta = \{P_\theta(x) : \theta \in \Theta \subset R^m\}$  on  $\mathcal{A}$ . Thus

$$P_\theta(X_1, \dots, X_N | N = n) = \prod_{i=1}^n P_\theta(X_i)$$

for  $n = 1, 2, \dots$ . Of particular interest is the class of all distributions with  $P_\theta(j) = \theta_j$  parameterized by the simplex  $\Theta = \{\theta = (\theta_1, \dots, \theta_m) : \theta_j \geq 0, \sum_{j=1}^m \theta_j = 1, j = 1, \dots, m\}$ .

Let  $\underline{N} = (N_1, \dots, N_m)$  denote the vector of counts for symbol  $1, \dots, m$ . The observed sample size  $N$  is the sum of the counts  $N = \sum_{j=1}^m N_j$ . Both  $P_\theta(\underline{X})$  and  $P_\theta(\underline{X} | N = n)$  have factorizations based on the distribution of the counts

$$P_\theta(\underline{X} | N = n) = P(\underline{X} | \underline{N}) P_\theta(\underline{N} | N = n),$$

and

$$P_\theta(\underline{X}) = P(\underline{X} | \underline{N}) P_\theta(\underline{N}).$$

The first factor of the two equations is the uniform distribution on the set of strings with given counts, which does not depend on  $\underline{\theta}$ . The vector of counts  $\underline{N}$  forms a sufficient statistic for  $\underline{\theta}$ . Modeling the distribution of the counts is essential for forming codes and predictions. In the particular case of all i.i.d. distributions parameterized by the simplex, the distribution  $P_{\underline{\theta}}(\underline{N}|N=n)$  is the *multinomial*( $n, \underline{\theta}$ ) distribution.

In the above, there is a need for a distribution of the total count  $N$ . Of particular interest is the case that the total count is taken to be *Poisson*, because then the resulting distribution of individual counts makes them independent.

Accordingly, we give particular attention to the target family  $\mathcal{P}_{\underline{\lambda}}^m = \{P_{\underline{\lambda}}(\underline{N}) : \lambda_j \geq 0, j=1, \dots, m\}$ , in which  $P_{\underline{\lambda}}(\underline{N})$  is the product of *Poisson*( $\lambda_j$ ) distribution for  $N_j, j=1, \dots, m$ . It makes the total count  $N \sim \text{Poisson}(\lambda_{sum})$  with  $\lambda_{sum} = \sum_{j=1}^m \lambda_j$  and yields the *multinomial*( $n, \underline{\theta}$ ) distribution by conditioning on  $N=n$ , where  $\theta_j = \lambda_j / \lambda_{sum}$ . And the induced distribution on  $\underline{X}$  is

$$P_{\underline{\lambda}}(\underline{X}) = P(\underline{X}|\underline{N})P_{\underline{\lambda}}(\underline{N}).$$

The task of coding a string is equivalent to providing a probabilistic scheme. A coder  $Q$  for the string is also a (sub)probability distribution on  $\mathcal{X}^*$  which assigns a probability  $Q(\underline{X})$  to each string  $\underline{X}$  and produces a binary string of length  $\log 1/Q(\underline{X})$  (we do not worry about the integer constraint). Ideally the true probability distribution  $P_{\underline{\lambda}}(\underline{X})$  could be used if  $\underline{\lambda}$  were known, as it produces no extra bits for coding purpose. The *regret* induced by using  $Q$  instead of  $P_{\underline{\lambda}}$  is

$$R(Q, P_{\underline{\lambda}}, \underline{X}) = \log \frac{1}{Q(\underline{X})} - \log \frac{1}{P_{\underline{\lambda}}(\underline{X})},$$

where  $\log$  is logarithm base 2. Likewise, the *expected regret* is

$$r(Q, P_{\underline{\lambda}}) = \mathbf{E}_{P_{\underline{\lambda}}} \left( \log \frac{1}{Q(\underline{X})} - \log \frac{1}{P_{\underline{\lambda}}(\underline{X})} \right).$$

In universal coding the expected regret is also called the *redundancy*.

Here we can construct  $Q$  by choosing a probability distribution for the counts and then use the uniform distribution for the distribution of strings given the counts, written as  $P_{unif}$ . That is

$$Q(\underline{X}) = P_{unif}(\underline{X}|\underline{N})Q(\underline{N}).$$

Then the regret becomes the log ratio of the counts probability

$$\begin{aligned} R(Q, P_{\underline{\lambda}}, \underline{X}) &= \log \frac{P_{\underline{\lambda}}(\underline{N})}{Q(\underline{N})} \\ &= R(Q, P_{\underline{\lambda}}, \underline{N}). \end{aligned}$$

And the redundancy becomes

$$r(Q, P_{\underline{\lambda}}) = \mathbf{E}_{P_{\underline{\lambda}}} \log \frac{P_{\underline{\lambda}}(\underline{N})}{Q(\underline{N})}.$$

In the pointwise regret story, the set of codelengths  $\log(1/P_{\underline{\lambda}}(\underline{X}))$  provides a standard with which our coder is to be compared. Given the family  $\mathcal{P}_{\underline{\lambda}}^m$ , consider the best candidate with hindsight  $P_{\hat{\underline{\lambda}}}(\underline{X})$ , which achieves the maximum value,  $P_{\hat{\underline{\lambda}}}(\underline{X}) = \max_{\underline{\lambda} \in \Lambda} (P_{\underline{\lambda}}(\underline{X}))$  (corresponding to  $\min_{\underline{\lambda} \in \Lambda} \log(1/P_{\underline{\lambda}}(\underline{X}))$ ), where  $\hat{\underline{\lambda}}$  is the maximum likelihood estimator of  $\underline{\lambda}$ , and compare it to our strategy  $Q(\underline{X})$ . The maximization is equivalent to maximizing  $\underline{\lambda}$  for the count probability, as the uniform distribution dose not depend on  $\underline{\lambda}$ , i.e.

$$\begin{aligned} \max_{\underline{\lambda} \in \Lambda} (P_{\underline{\lambda}}(\underline{X})) &= P_{unif}(\underline{X}|\underline{N}) \max_{\underline{\lambda} \in \Lambda} P_{\underline{\lambda}}(\underline{N}) \\ &= P_{unif}(\underline{X}|\underline{N}) P_{\hat{\underline{\lambda}}}(\underline{N}). \end{aligned}$$

Then the problem becomes: given the family  $\mathcal{P}_{\underline{\lambda}}^m$ , how to choose  $Q$  to minimize the maximized regret

$$\min_Q \max_{\underline{X}} R(Q, P_{\underline{\lambda}}, \underline{X}) = \min_Q \max_{\underline{N}} \log \frac{P_{\hat{\underline{\lambda}}}(\underline{N})}{Q(\underline{N})},$$

or the redundancy,

$$\min_Q \max_{\underline{P}_{\underline{\lambda}} \in \mathcal{P}_{\underline{\lambda}}^m} r(Q, P_{\underline{\lambda}}) = \min_Q \max_{\underline{P}_{\underline{\lambda}} \in \mathcal{P}_{\underline{\lambda}}^m} \mathbf{E}_{P_{\underline{\lambda}}} \log \frac{P_{\hat{\underline{\lambda}}}(\underline{N})}{Q(\underline{N})}.$$

For the regret, the maximum can be restricted to a set of counts instead of the whole space. A traditional choice being  $S_{m,n} = \{(N_1, \dots, N_m) : \sum_{j=1}^m N_j = n, N_j \geq 0, j=1, \dots, m\}$  associated with a given sample size  $n$ , in which case the minimax regret is

$$\min_Q \max_{\underline{N} \in S_{m,n}} \log \frac{P_{\hat{\underline{\lambda}}}(\underline{N})}{Q(\underline{N})}.$$

As is familiar in universal coding [2][3], the normalized maximum likelihood (NML) distribution

$$Q_{nml}(\underline{N}) = \frac{P_{\hat{\lambda}}(\underline{N})}{C(S_{m,n})}$$

is the unique pointwise minimax strategy when  $C(S_{m,n}) = \sum_{\underline{N} \in S_{m,n}} P_{\hat{\lambda}}(\underline{N})$  is finite, and  $\log C(S_{m,n})$  is the minimax value. When  $m$  is large, the NML distribution can be unwieldy to compute for compression or prediction. Instead we will introduce a slightly suboptimal coding distribution that makes the counts independent and show that it is nearly optimal for every  $S_{m,n'}$  with  $n'$  not too different from a target  $n$ . Indeed, we advocate that our simple coding distribution is preferable to use computationally when  $m$  is large even if the sample size  $n$  were known in advance.

To produce our desired coding distribution we make use of two basic principles. One is that the multinomial family of distributions on counts matches the conditional distribution of  $N_1, \dots, N_m$  given the sum  $N$  when unconditionally the counts are independent Poisson. Another is the information theory principle [4][5][6] that the conditional distribution given a sum (or average) of a large number of independent random variables is approximately a product of distributions, each of which is the one closest in relative entropy to the unconditional distribution subject to an expectation constraint. This minimum relative entropy distribution is an exponential tilting of the unconditional distribution.

In the Poisson family with distribution  $\lambda_j^{N_j} e^{-\lambda_j} / N_j!$ , exponential tilting (multiplying by the factor  $e^{-a N_j}$ ) preserves the Poisson family (with the parameter scaled to  $\lambda_j e^{-a}$ ). Those distributions continue to correspond to the multinomial distribution (with parameters  $\theta_j = \lambda_j / \lambda_{sum}$ ) when conditioning on the sum of counts  $N$ . A particular choice of  $a = \ln(\lambda_{sum} / N)$  provides the product of Poisson distributions closest to the multinomial in regret. Here for universal coding, we find the tilting of individual maximized likelihood that makes the product of such closest to the Shtarkov's NML distribution. This greatly simplifies the task of approximate optimal universal compression and the analysis of its regret.

Indeed, applying the maximum likelihood step to a Poisson count  $k$  produces a maximized likelihood value of  $M(k) = k^k e^{-k} / k!$ . We call this maximized likelihood the Stirling ratio, as it is the quantity that Stirling's approximation shows near  $(2\pi k)^{-1/2}$  for  $k$  not small. We find that this  $M(k)$  plays a distinguished role in universal large alphabet compression, even for sequences with small counts  $k$ . This measure  $M$  has a product extension to counts  $N_1, N_2, \dots, N_m$ ,

$$M(\underline{N}) = M(N_1)M(N_2) \cdots M(N_m).$$

Although  $M$  has an infinite sum by itself, it is normalizable when tilted for every positive  $a$ . The tilted Stirling ratio distribution is

$$P_a(N_j) = \frac{N_j^{N_j} e^{-N_j}}{N_j!} \frac{e^{-a N_j}}{C_a}, \quad (1)$$

with the normalizer  $C_a = \sum_{k=0}^{\infty} k^k e^{-(1+a)k} / k!$ .

The coding distribution we propose and analyze is simply the product of those tilted one-dimensional maximized Poisson likelihood distributions for a value of  $a$  we will specify later

$$Q_a(\underline{N}) = P_a^m(\underline{N}) = P_a(N_1) \cdots P_a(N_m).$$

By allowing description of all possible counts  $N_j \geq 0$ ,  $j = 1, \dots, m$ , our codelength will be greater for some strings than codelengths designed for the case of a given sum  $N = n$ . Nevertheless, with  $N$  distributed  $Poisson(n)$ , the probability of the outcome  $N = n$  is approximately  $P(N = n) \approx 1/\sqrt{2\pi n}$ . So the allowance of description of  $N$  (not just  $N_1, \dots, N_m$  given  $N$ ) adds  $\log 1/P(N = n)$  which is approximately  $\frac{1}{2} \log 2\pi n$  bits to the description length beyond that which would have been ideal  $\log 1/Q_a(N_1, \dots, N_m | N = n)$  if  $N = n$  were known. This ideal codelength constructed from the tilted maximized Poisson, when conditioning on  $n$ , matches the Shtarkov's normalized maximum likelihood based on the multinomial.

For small alphabet with  $m \ll n$ , the minimax regret is about  $\frac{1}{2} \log n$  bits per free parameter (a total of  $\frac{m-1}{2} \log n + \text{constant}$ ); and for large alphabet when  $m \sim n$  and  $n = o(m)$ , the minimax regret is about  $O(n)$  and  $n \log \frac{m}{n}$  respectively [2][3][7][8]. The additional  $\frac{1}{2} \log n$  bits is a small price to pay for the sake of gaining the coding simplification and additional flexibility.

If it is known that the total count is  $n$ , then the regret is a simple function of  $n$  and the normalizer  $C_a$ . The choice of the tilting parameter  $a^*$  given by the moment condition  $\mathbf{E}_{Q_a} \sum_{j=1}^m N_j = n$  minimizes the regret over all positive  $a$ . This arises by differentiation because  $\frac{\partial}{\partial a} \log C_a$  is equal to the given moment. Moreover,  $a^*$  depends only on the ratio between the size of the alphabet and the total count  $m/n$ . Fig. 1 displays  $a^*$  as a function of  $m/n$  solved numerically. Given an alphabet with  $m$  symbols and a string generated from it of length  $n$ , one can look at the plot and find the  $a^*$  desired according to the  $m/n$  given, and then use the  $a^*$  to code the data.

If, however, the total count  $N$  is not given, then the regret depends on  $N$ . We use a mixture of  $a$  to account for the lack of knowledge in advance, and details are discussed in section III-D.

When  $a$  is small, the tilting of the maximized Poisson likelihood distributions does not have much effect except in the tail of the distribution. Over most of the range of count values  $k$  it follows the approximate power-law  $1/k^{1/2}$  as we have indicated. Power-laws have been studied for count distributions and are shown to be related to Zipf's law for the sorted counts [9]. Our

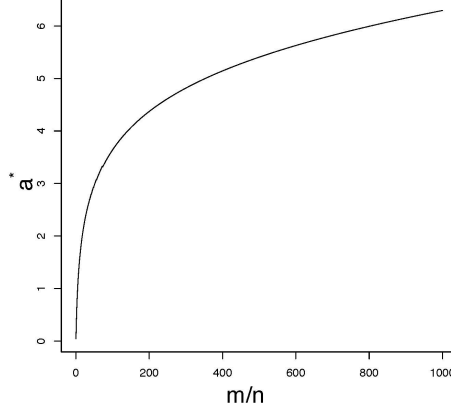


Fig. 1. Relationship between  $a^*$  and  $\frac{m}{n}$ .

use of a distribution close to a power-law is not because a power-law is assumed to govern the data, but rather because of its near optimum regret properties within suitable sets of counts, demonstrated here for the class of all Poisson count distributions, from which we obtain also its near optimality for the class of all Multinomial distributions on counts.

Shtarkov studied the universal data compression problem and identified the exact pointwise minimax strategy [2]. He showed the asymptotic minimax lower bound for the regret is  $\frac{m-1}{2} \log n + O(1)$ , in which the parameter set  $\Theta$  is the  $m-1$  dimensional simplex of all probability vectors on an alphabet of size  $m$ . However, this strategy cannot be easily implemented for prediction or compression [2], because of the computational inconvenience of computing the normalizing constant, and because of the difficulty in computing the successive conditionals required for implementation (by arithmetic coding). Let  $m^*$  be the number of different symbols that appear in a sequence. Shtarkov[10] also pointed out that when  $m$  is large, it typical that  $m^*$  is much less than  $m$ , and the regret depends mainly on  $m^*$  rather than  $m$ . Xie and Barron[3][11] gave an asymptotic minimax strategy for coding under both the expected and pointwise regret for fixed size alphabet, which is formulated by a modification of the mixture density using Jeffery's prior. The asymptotic value of both the redundancy and the regret are of the form  $\frac{m-1}{2} \log n + C_m + o(1)$ , where  $C_m$  is a constant depending on  $m$ . Orlitsky and Santhanam[12] considered the problem in a large alphabet setting in which the number of symbols  $m$  is much larger than the sequence length  $n$  or even infinite. They found the main terms in the minimax regret for  $m = o(n)$ ,  $m \sim n$  and  $n = o(m)$  cases take the forms  $\frac{m-1}{2} \log \frac{n}{m}$ ,  $O(m)$  and  $n \log \frac{m}{n}$  respectively. Szpankowski and Weinberger[8] provided more precise asymptotics in these settings. They also calculated the minimax regret of a source model in which some symbol probabilities are fixed. Boucheron, Garivier and Gassiat[13] focused on countably infinite alphabets with an envelope condition; they used an adapted strategy and gave upper and lower bounds for pointwise minimax regret. Later on Bontemps and Gassiat[14] worked on exponentially decreasing envelope class and provided a minimax strategy and the corresponding regret.

In this paper, we introduce a straightforward and easy to implement method for large alphabet coding. The purpose is three-fold: first, by allowing the sample size to be variable, we are considering a larger class of distributions. This is a more realistic and less restrictive assumption than presuming a particular length. But the method can also be used for fixed sample size coding and prediction.

Second, it unveils an information geometry of three key distributions/measures in the problem: the unnormalized maximum Poisson likelihood measure  $M$  of the counts, the conditional distribution  $M_{cond}$  of  $M$  given the total count equals  $n$ , which matches Shtarkov's normalized maximum multinomial likelihood distribution, and a tilted distribution  $Q_a$ , with the tilting parameter  $a$  chosen to make the expected total count equal to  $n$ . This tilted distribution  $Q_a$  minimizes the relative entropy from the original measure  $M$  within the class  $\mathcal{C}$  of distributions with the moment condition  $E[N] = n$ . Hence,  $Q_a$  is the information projection of  $M$  onto  $\mathcal{C}$ . Moreover, since  $M_{cond}$  is also in  $\mathcal{C}$ , the Pythagorean-like equality holds [15][4], i.e.

$$D(M_{cond}||M) = D(M_{cond}||Q_a) + D(Q_a||M).$$

The case of a tilted distribution (the information projection) as an approximating conditional distribution is investigated in [6] and [5]. A difference here is that our unconditional measure  $M$  is not normalizable.

Thirdly, the strategy designed through an independent Poisson model and tilting is much easier to analyze and compute as compared to the strategies based on multinomials. The convenience is gained through independence. To actually apply this two pass code, one could first describe the independent counts  $N_1, \dots, N_m$ , for instance by arithmetic coding using  $P_a(N_j)$ , and then describe  $X_1, \dots, X_n$  given the count vector, by arithmetic coding using the sequence of conditional distributions for

$X_{i+1}$  given both  $X_1, \dots, X_i$  and all the counts (which is the sampling without replacement distribution, proportional to the counts of what remains after step  $i$ ).

This paper is organized in the following way. Section II introduces the model. Section III provides general results and outlines the proof, and Section IV gives simulated and real data examples. Details of proof are left in the appendix.

## II. THE POISSON MODEL

A Poisson model fits well into this problem. We have for each  $j = 1, \dots, m$ ,

$$N_j \sim \text{Poisson}(\lambda_j),$$

independently, and  $N$  also has a Poisson distribution

$$N \sim \text{Poisson}(\lambda_{\text{sum}}),$$

where  $\lambda_{\text{sum}} = \sum_{j=1}^m \lambda_j$ . Write  $\underline{\lambda} = (\lambda_1, \dots, \lambda_m)$ , we have

$$P_{\underline{\lambda}}(\underline{X}) = P_{\text{unif}}(\underline{X}|\underline{N}) \prod_{j=1}^m P_{\lambda_j}(N_j).$$

We know that the the MLE for each  $\lambda_j$  is  $\hat{\lambda}_j = N_j$ , and the first term is a uniform distribution which does not depend on  $\underline{\lambda}$ . So

$$P_{\hat{\underline{\lambda}}}(\underline{X}) = P_{\text{unif}}(\underline{X}|\underline{N}) \prod_{j=1}^m M(N_j).$$

where  $M(k) = k^k e^{-k} / k!$ ,  $k = 1, 2, \dots$  (as given in the introduction) is the unnormalized maximized likelihood  $M(N_j) = \max_{\lambda_j} P_{\lambda_j}(N_j)$ .

If we use a distribution  $Q(\underline{N})$  to code the counts, then the regret is

$$\log \frac{P_{\hat{\underline{\lambda}}}(\underline{X})}{P(\underline{X}|\underline{N})Q(\underline{N})} = \log \frac{\prod_{j=1}^m M(N_j)}{Q(\underline{N})}.$$

And the redundancy is

$$\mathbf{E}_{P_{\underline{\lambda}}} \log \frac{P(\underline{X}|\underline{\lambda})}{P(\underline{X}|\underline{N})Q(\underline{N})} = \mathbf{E}_{P_{\underline{\lambda}}} \log \frac{P(\underline{N}|\underline{\lambda})}{Q(\underline{N})}.$$

This method can also be applied to fixed total count scenario, which corresponds to the multinomial coding and prediction problem. Suppose  $N = n$  is given, the Poisson model, when conditioned on  $N = n$ , indeed reduces to the i.i.d sampling model

$$P_{\underline{\lambda}}(X_1, \dots, X_N | N = n) = P_{\underline{\theta}}(X_1, \dots, X_n).$$

The right hand side is a discrete memoryless source distribution (i.i.d.  $P_{\underline{\theta}}$ ) with probability specified by  $P_{\underline{\theta}}(j) = \theta_j$ , for  $j = 1, \dots, m$ . Note that a sequence  $X_1, \dots, X_N$  with counts  $N_1, \dots, N_m$  of total  $N = n$  satisfies

$$\begin{aligned} & P_{\underline{\lambda}}(X_1, \dots, X_N | N = n) \\ &= \frac{P_{\underline{\lambda}}(X_1, \dots, X_n)}{P_{\lambda_{\text{sum}}}(N = n)} \\ &= \frac{P_{\text{unif}}(X_1, \dots, X_n | N_1, \dots, N_m) P_{\underline{\lambda}}(N_1, \dots, N_m)}{P_{\lambda_{\text{sum}}}(N = n)}. \end{aligned}$$

The question left is still how to model the counts. The maximized likelihood (the same target as used by Shtarkov) is thus expressible as

$$\begin{aligned} & P_{\hat{\underline{\lambda}}}(X_1, \dots, X_N | N = n) \\ &= \frac{P_{\text{unif}}(X_1, \dots, X_n | N_1, \dots, N_m) \prod_{j=1}^m M(N_j)}{P_{\hat{\lambda}_{\text{sum}}}(N = n)}. \end{aligned}$$

Now again if we use  $Q(N_1, \dots, N_m)$  to code the counts, then the regret is

$$\begin{aligned} & \log \frac{P_{\hat{\underline{\lambda}}}(X_1, \dots, X_N | N = n)}{P_{\text{unif}}(X_1, \dots, X_n | N_1, \dots, N_m) Q(N_1, \dots, N_m)} \\ &= \log \frac{\prod_{j=1}^m M(N_j)}{P_{\hat{\lambda}_{\text{sum}}}(N = n) Q(N_1, \dots, N_m)} \\ &\simeq \frac{1}{2} \log 2\pi n + \log \frac{\prod_{j=1}^m M(N_j)}{Q(N_1, \dots, N_m)} \end{aligned} \tag{2}$$

Here  $\hat{\lambda}_{sum} = n$ , hence the term  $\frac{1}{2} \log 2\pi n$  is Stirling's approximation of  $\log 1/P_{\hat{\lambda}_{sum}}(N = n)$ . The  $\frac{1}{2} \log 2\pi n$  arises because here  $Q$  includes description of the total  $N$  while the more restrictive target regards it as given.

### III. RESULTS

#### A. Regret

We start by looking at the performance of using independent tilted Stirling ratio distributions as a coding strategy, by examining the resulting regret.

Let  $S$  be any set of counts, then the maximized regret of using  $Q$  as a coding strategy given a class  $\mathcal{P}$  of distributions when the vector of counts is restricted to  $S$  is

$$R(Q, \mathcal{P}, S) = \max_{\underline{N} \in S} \log \frac{\max_{P \in \mathcal{P}} P(\underline{N})}{Q(\underline{N})}.$$

**Theorem 1.** Let  $P_a$  be the distribution specified in equation (1) (Poisson maximized likelihood, tilted and normalized). The regret of using a product of tilted distributions  $Q_a = \otimes_{j=1}^m P_a$  for a given vector of counts  $\underline{N} = (N_1, \dots, N_m)$  is

$$R(Q_a, \mathcal{P}_\Lambda^m, \underline{N}) = aN \log e + m \log C_a.$$

Let  $S_{m,n}$  be the set of count vectors with total count  $n$  be defined as before, then

$$R(Q_a, \mathcal{P}_\Lambda^m, S_{m,n}) = an \log e + m \log C_a. \quad (3)$$

Let  $a^*$  be the choice of  $a$  satisfying the following moment condition

$$\mathbf{E}_{P_a} \sum_{j=1}^m N_j = m \mathbf{E}_{P_a} N_1 = n. \quad (4)$$

Then  $a^*$  is the minimizer of the regret in expression (3). Write  $R_{m,n} = \min_a R(Q_a, \mathcal{P}_\Lambda^m, S_{m,n})$ .

When  $m = o(n)$ , the  $R_{m,n}$  is near  $\frac{m}{2} \log \frac{ne}{m}$  in the following sense.

$$\begin{aligned} -d_1 \frac{m}{2} \log e &\leq R_{m,n} - \frac{m}{2} \log \frac{ne}{m} \\ &\leq m \log \left(1 + \sqrt{\frac{m}{n}}\right), \end{aligned} \quad (5)$$

where  $d_1 = O\left(\left(\frac{m}{n}\right)^{1/3}\right)$ .

When  $n = o(m)$ , the  $R_{m,n}$  is near  $n \log \frac{m}{ne}$  in the following sense.

$$\begin{aligned} m \log \left(1 + (1 - d_2) \frac{n}{m}\right) &\leq R_{m,n} - n \log \frac{m}{ne} \\ &\leq m \log \left(1 + \frac{n}{m} + d_3\right) \end{aligned} \quad (6)$$

where  $d_2 = O\left(\frac{n}{m}\right)$ , and  $d_3 = \frac{1}{2\sqrt{\pi}} \frac{n^2 e^2}{m(m - ne)}$ .

When  $n = bm$ , the  $R_{m,n} = cm$ , where the constant  $c = a^* b \log e + \log C_{a^*}$ , and  $a^*$  is such that  $\mathbf{E}_{P_a} N_1 = b$ .

*Proof:* The expression of the regret is from the definition. The fact that  $a^*$  is the minimizer can be seen by taking partial derivative with respect to  $a$  of expression (3). The upper bounds are derived by applying Lemma 1 in the appendix. Pick  $a = m/2n$  and use the first inequality, we get the upper bound for  $m = o(n)$  case; pick  $a = \ln(m/ne)$  and use the second inequality, we have the upper bound for  $n = o(m)$ . Here  $\ln$  is the logarithm base  $e$ . The rest of the proof is left in Appendix B. ■

**Remark 1:** The regret depends only on the number of parameters  $m$ , the total counts  $n$  and the tilting parameter  $a$ . The optimal tilting parameter is given by a simple moment condition in equation (4).

**Remark 2:** The regret  $R_{m,n}$  is close to the minimax level in all three cases listed in Theorem 1. The main terms in the  $m = o(n)$  and  $n = o(m)$  cases are the same as the minimax regret given in [8] except the multiplier for  $\log(ne/m)$  here is  $m/2$  instead of  $(m-1)/2$  for the small  $m$  scenario. For the  $n = bm$  case, the  $R_{m,n}$  is close to the minimax regret in [8] numerically.

**Remark 3:** In fact, the regret provides an upper bound for the regret. Recall that

$$\begin{aligned} \mathbf{E}_{P_\Lambda} \log \frac{P_\Lambda}{Q_a} &\leq \mathbf{E}_{P_\Lambda} \max_{\underline{\lambda}} \log \frac{P_\Lambda}{Q_a} \\ &= a \lambda_{sum} \log e + m \log C_a. \end{aligned} \quad (7)$$

Theorem 4 in Appendix C gives more detailed expression of the redundancy for using  $Q_a$ . While there is a reduction of  $(m/2) \log e$  bits as compared to the pointwise case, the error depends on the  $\lambda_j$ 's. Nevertheless, expression (7) still provides an uniform upper bound for the redundancy for all possible Poisson means  $\underline{\lambda}$  with a given sum.

**Corollary 1.** Let  $\mathcal{P}_\Theta^m$  be a family of multinomial distributions with total count  $n$ . Then the maximized regret  $R(Q_a, \mathcal{P}_\Theta^m, S_{m,n})$  has an upper bound within  $\frac{1}{2} \log 2\pi n$  above the upper bound in Theorem 1.

*Proof:* This can be easily seen by equation (2). ■

### B. Subset of sequences with partitioned counts

One advantage of using the tilted Stirling ratio distributions is the flexibility of choosing tilting parameters. As mentioned in the introduction, the ratio  $m/n$  uniquely determines the optimal tilting parameter. In fact, different tilting parameters can be used for symbols to adjust for their relative importance in the alphabet. Here we consider a situation in which the empirical distribution has most probability captured by a small portion of the symbols. This happens when the sorted probability list is quite skewed.

The following theorem holds for strings with constraints on the sum of tail counts  $\sum_{j>L} N_j = nf$ . Small remainder occurs in the following regret bound when  $nf/(m-L)$  and  $L/(n-nf)$  are both small.

**Theorem 2.** Let  $S_{m,n,f,L}$  be a subset of count vectors with the tail sum controlled by a value  $0 \leq f \leq 1$ , that is,  $S_{m,n,f,L} = \{\underline{N} = (N_1, \dots, N_m) : \sum_{j=1}^m N_j = n, \sum_{j>L} N_j = nf\}$ . Here  $L$  is a number between 0 and  $m$ . The regret of using the tilted Stirling ratio distributions for count vectors in  $S_{m,n,f,L}$  given each  $L \in \{0, \dots, m\}$  is mainly

$$\frac{L}{2} \log \frac{(n-nf)e}{L} + nf \log \frac{(m-L)}{nfe}. \quad (8)$$

The remainder is bounded below by  $r_1$  and above by  $r_2$ , where

$$r_1 = -d_1 \frac{L}{2} \log e + (m-L) \log \left( 1 + (1-d_2) \frac{nf}{m-L} \right),$$

and

$$\begin{aligned} r_2 &= (m-L) \log \left( 1 + \frac{nf}{m-L} + d_3 \right) \\ &\quad + L \log \left( 1 + \sqrt{\frac{L}{n-nf}} \right). \end{aligned}$$

Here  $d_1$  is  $O\left(\left(\frac{L}{n-nf}\right)^{1/3}\right)$  and  $d_2$  is  $O\left(\frac{nf}{m-L}\right)$  and  $d_3 = \frac{1}{2\sqrt{\pi}} \frac{(nfe)^2}{(m-L)((m-L)-nfe)}$ .

*Proof:* Consider the product distribution,

$$\begin{aligned} Q_{a,b}(\underline{N}) &= \prod_{j=1}^m P_{a,b}(N_j) \\ &= \prod_{j=1}^m \frac{N_j^{N_j} e^{-N_j}}{N_j!} \frac{e^{-aN_j} e^{-bN_j} \mathbf{1}_{\{j>L\}}}{C_{a,b,j}}, \end{aligned}$$

where  $C_{a,b,j} = C_a$  if  $j \leq L$ , and  $C_{a,b,j} = C_{a,b}$  is defined as  $\sum_{k=0}^{\infty} k^k e^{-(1+a+b)k} / k!$  if  $j > L$ . It is in fact using an  $L$  dimensional product distribution  $Q_a$  on the first  $L$  symbols, and an  $m-L$  dimensional product distribution  $Q_{a+b}$  on the rest.

The regret is the same for any  $\underline{N} \in S_{m,n,f,L}$  given  $a$  and  $b$ . That is,

$$\begin{aligned} R(Q_{a,b}, \mathcal{P}_\Lambda^m, S_{m,n,f,L}) &= n \log e + L \log C_a + nfb \log e + (m-L) \log C_{a,b} \\ &= R(Q_a, \mathcal{P}_\Lambda^L, S_{L,n-nf}) + R(Q_{a+b}, \mathcal{P}_\Lambda^{m-L}, S_{m-L,nf}). \end{aligned}$$

Here  $\mathcal{P}_\Lambda^j$  denotes the class of  $j$  independent Poisson distributions and  $S_{j,k}$  is the set of  $j$  independent Poisson counts with sum equal to  $k$ . In the above case,  $j = L$  or  $m-L$ , and  $k = n-nf$  or  $nf$ .

The choice of  $a, b$  providing minimization of  $R(Q_{a,b}, \mathcal{P}_\Lambda^m, S_{m,n,f,L})$  is given by the following conditions

$$\begin{aligned} \mathbf{E}_{P_{a,b}} \sum_{j=1}^m N_j &= n \\ \mathbf{E}_{P_{a,b}} \sum_{j>L} N_j &= nf. \end{aligned}$$

This result can be derived by applying Theorem 1 to  $R(Q_a, \mathcal{P}_\Lambda^L, S_{L,n-nf})$  and  $R(Q_{a+b}, \mathcal{P}_\Lambda^{m-L}, S_{m-L,nf})$  respectively. ■

**Remark 4:** The problem here is treated as two separate coding tasks, one for a small alphabet with  $L$  symbols having a total count  $n - nf$ , and the other for a large alphabet with  $m - L$  symbols with total count  $nf$ . The two main terms in expression (8) represent regret from coding the two subsets of symbols, with one set containing  $L$  symbols having relatively large counts, and each symbol induces  $\frac{1}{2} \log \frac{n(1-f)e}{L}$  bits of regret, and the other containing the rest  $m - L$  symbols with small counts and together cost  $nf \log \frac{m}{nfe}$  extra bits.

**Remark 5:** One can arrange more flexibility in what the code can achieve by adding small additional pieces to the code. One is to adapt the choice of  $L$  between 0 and  $m$ , including  $\log(m+1)$  more bits for the description of  $L$ . Next one can either work with the counts in the given order, or use an additional  $\log \binom{m}{L}$  bits to describe the subset that has the  $L$  largest counts. Then one uses  $\log 1/Q_{a,b}(\underline{N})$  bits to describe the counts. Rather than fixing  $f$ , one works with the empirical tail fraction  $\hat{f}(L)$ , where  $n\hat{f}(L)$  is the sum of the counts for the remaining  $m - L$  symbols. Finally one has to adapt the choices of  $a$  and  $b$ . A suggested method of doing so is described in Section III-D, in which the  $Q_{a,b}$  above is replaced by a mixture over a range of choices of  $a$  and  $b$ .

### C. Envelope class

Besides a subset of strings, we can also consider subclass of distributions. Here we follow the definition of envelope class in [13]. Suppose  $\mathcal{P}_{m,f}$  is a class of distributions on  $1, \dots, m$  with the symbol probability bounded above by an envelope function  $f$ , i.e.

$$\mathcal{P}_{m,f} = \{P_\theta : \theta_j \leq f(j), j = 1, \dots, m\}.$$

Given the string length  $n$ , we know the count of each symbol follows a Poisson distribution with mean  $\lambda_j = n\theta_j$ ,  $j = 1, \dots, m$ . This transfers an envelope condition from the multinomial distribution to a Poisson distribution, the mean for which is restricted to the following set

$$\Lambda_{m,f} = \{\lambda : \lambda_j \leq nf(j), j = 1, \dots, m\}.$$

**Theorem 3.** *The minimax regret of the Poisson class  $\Lambda_{m,f}$  with envelope function  $f$  has the following upper bound*

$$\begin{aligned} & R(Q_a, \Lambda_{m,f}, \underline{N}) \\ & \leq \min_{L \in \{1, \dots, m\}} \frac{L}{2} \log \frac{n(1 - \bar{F}(L))}{L} + n\bar{F}(L) \log e + r_3, \end{aligned}$$

where  $\bar{F}(L) = \sum_{j>L} f(j)$ , and

$$r_3 = \frac{L}{2(1 - \bar{F}(L))} \log e + L \log \left( 1 + \sqrt{\frac{L}{n(1 - \bar{F}(L))}} \right).$$

*Proof:* A tilted distribution with  $a = L/2n(1 - \bar{F}(L))$  will give the result. Details are left in Appendix D. ■

**Remark 6:** Here in order for  $r_3$  to be small, the tail sum of the envelope function  $\bar{F}(L)$  needs to be small, although the upper bound holds for general envelope function  $f$  and  $L$ . This result is of the same order as the upper bound  $\inf_{L: L \leq n} ((L-1)/2 \log n + n\bar{F}(L))$  2 given in [13]. The first main term in the bound given in Theorem 3 also matches the minimax regret given in [3] for an alphabet with  $L$  symbols and  $n(1 - \bar{F}(L))$  data points by Stirling's approximation, i.e.,

$$\begin{aligned} & \frac{L-1}{2} \log \frac{n(1 - \bar{F}(L))}{2\pi} + \log \frac{\Gamma(1/2)^L}{\Gamma(L/2)} \\ & \approx \frac{L-1}{2} \log \frac{n(1 - \bar{F}(L))e}{L} + \frac{1}{2} \log \frac{e}{2}. \end{aligned}$$

The extra  $(1/2) \log(n(1 - f)e/L)$  is because the tilted distribution allows  $m$  free parameters instead of  $m - 1$ .

**Remark 7:** The best choice of tilting parameters for envelope class only depends on the envelope function and the number of symbols  $L$  constituting the ‘frequent’ subset. Unlike the subset of strings case discussed before, neither the order of the counts nor which symbols are those with largest counts matters, all we need is an envelope function decaying fast enough when the symbol probabilities are arranged in decreasing order so that  $L$  and  $\bar{F}(L)$  are both small.



#### D. Regret with unknown total count

We know that  $a^*$  depends on the value of the total count. However, when the total count is not known, we can use a mixture of tilted distributions  $Q(\underline{N})$ .

$$\begin{aligned} Q(\underline{N}) &= \int_0^{m/2} Q_a(\underline{N}) \frac{1}{m/2} da \\ &= \int_0^{m/2} \prod_{j=1}^m \frac{N_j^{N_j} e^{-N_j}}{N_j! C_a} e^{-aN_j} \frac{2}{m} da \\ &\leq M(\underline{N}) \frac{2}{m} \int_0^\infty e^{-aN} C_a^{-m} da \end{aligned}$$

Here the upper end of the integrated area is due to inequality (16). We have  $a^* \leq m/(2n) \leq m/2$ .

For any realized non-negative total count  $N = k$ , the integrand is maximized at  $a_k^*$ , defined as solution to the equation  $\mathbf{E}_{P_a} N = k$ . And the integral can be approximated by Laplace method,

$$Q(\underline{N}) \approx \frac{2}{m} \left( \prod_{j=1}^m \frac{N_j^{N_j} e^{-N_j}}{N_j!} \right) e^{-a_k^* k} C_{a_k^*}^{-m} \sqrt{\frac{2\pi}{c}},$$

where  $c = -\frac{\partial^2}{\partial a^2} \ln(e^{-ak} C_a^{-m})|_{a=a_k^*}$ .

Hence the regret induced by  $Q(\underline{N})$  is

$$\log \frac{M(\underline{N})}{Q(\underline{N})} \approx \log e^{a_k^* k} C_{a_k^*}^m + \log \sqrt{\frac{c}{2\pi}} + \log \frac{m}{2}.$$

By definition,

$$\begin{aligned} c &= m \left( \frac{\frac{\partial^2 C_a}{\partial a^2}}{C_a} - \left( \frac{\frac{\partial C_a}{\partial a}}{C_a} \right)^2 \right) \Big|_{a=a_k^*} \\ &\leq m \left( \frac{\frac{\partial^2 C_a}{\partial a^2}}{C_a} \right) \Big|_{a=a_k^*}. \end{aligned}$$

A similar argument as in the proof of Lemma 2 yields an upper bound for the first term

$$\begin{aligned} \frac{\frac{\partial^2 C_a}{\partial a^2}}{C_a} &\leq \frac{3}{(2a)^2} + \frac{1}{\sqrt{2\pi}} \left( \frac{3}{e} \right)^{3/2} \frac{1}{2a} \\ &\leq 3C_a^4 + \frac{1}{\sqrt{2\pi}} \left( \frac{3}{e} \right)^{3/2} C_a^2 \\ &< \frac{7}{2} C_a^4. \end{aligned}$$

The second last inequality is by Lemma 1.

Hence, we have an upper bound for the regret

$$\begin{aligned} &\log e^{a_k^* k} C_{a_k^*}^m + \log \frac{m}{2} + \frac{1}{2} \log \left( \frac{7m}{4\pi} C_{a_k^*}^4 \right) \\ &< \log e^{a_k^* k} C_{a_k^*}^m + \frac{3}{2} \log m + 2 \log C_{a_k^*}. \end{aligned}$$

Thus, the extra regret above the optimal level by using  $Q(\underline{N})$  is approximately no more than  $\frac{3}{2} \log m + 2 \log C_{a_k^*}$  bits.

Similar argument can show that averaging over the two parameters tilting distribution  $Q_{a,b}$  can lead to a distribution that achieves regret not much larger than the minimizing value if the actual total count and tail sum were known beforehand.

#### E. Conditional distributions induced by $Q_a(\underline{N})$

To account for strings of arbitrary length, our coding strategy  $Q_a$  assigns a probability distribution to all finite length strings on  $\mathcal{A}^m$ . However, when considering strings of a known length, we are interested to see what the distribution looks like conditioning on a particular number  $n$ .

Let  $\underline{N}^n$  denote any count vector in  $S_{m,n}$ , and  $N_x^n$  denote the  $x$ 's component of  $\underline{N}^n$ , where  $x \in \{1, \dots, m\}$ . Also, let  $M_{mul}$  be the *multinomial*( $n, \underline{\theta}$ ) maximized likelihood. We have

$$Q_a(\underline{N}^n | N = n) = \frac{Q_a(\underline{N}^n)}{Q_a(S_{m,n})} = \frac{M_{mul}(\underline{N}^n)}{M_{mul}(S_{m,n})}. \quad (9)$$

The conditioning of  $Q_a$  in expression (9) reduces the Poisson maximized likelihood (conditioned on the sum  $N = n$ ) to be the same as the multinomial maximized likelihood normalized as indicated, which is indeed the Shtarkov NML distribution for the multinomial family of distributions of counts.

This conditional distribution of counts, when multiplied by the uniform distribution of strings given the counts, induces a distribution on the strings, i.e.,

$$P_n(\underline{X}^n) = P_{unif}(\underline{X}^n | \underline{N}^n) Q_a(\underline{N}^n | N = n),$$

where  $\underline{X}^n$  is the vector  $X_1, \dots, X_n$ .

This sequence of distributions  $P_n$  are not compatible and hence do not have extensions to a stochastic process. To see this incompatibility one looks at the sum

$$\sum_{x \in \mathcal{A}} P_{n+1}(X_1, \dots, X_n, X_{n+1} = x)$$

and confirms that it is not equal to  $P_n(X^n)$ . This property is what is called the horizon dependence of NML [16].

### F. Prediction

A sequence of conditional distributions for  $X_{i+1}$  given the past observations  $X_1, \dots, X_i$  for  $i < n$  provides a sequential prediction with cumulative log loss defined by  $\sum_{i < n} \log 1/P(X_{i+1} | X_1, \dots, X_i)$ .

There are two natural ways of providing this sequence of conditionals. One is to get the conditionals from the full joint distribution  $P_n$ , which is horizon dependent as mentioned above. It produces cumulative log loss prediction regret precisely the same as the regret of using  $Q_a$  for data compression. The other is by using the sequence of distributions  $P_{i+1}(X_1, \dots, X_{i+1})$ ,  $i < n$ , called sequential NML [17]. The sequential prediction distribution  $P_{i+1}(X_{i+1} = x | X_1, \dots, X_i)$  is proportional to  $P_{i+1}(X_1, \dots, X_i, X_{i+1} = x)$  and accordingly simplifies to

$$P(X_{i+1} = x | X_1, \dots, X_i) = \frac{(N_x^i + 1)^{N_x^i + 1} / N_x^{i N_x^i}}{\sum_{\tilde{x}=1}^m (N_{\tilde{x}}^i + 1)^{N_{\tilde{x}}^i + 1} / N_{\tilde{x}}^{i N_{\tilde{x}}^i}}.$$

Note that the prediction rule does not involve  $a$ . Previous study by Shtarkov[2] shows that it is approximately proportional for large  $N_x$  to the  $N_x + 1/2$  rule of the Jeffreys *Beta*( $1/2, 1/2$ ) mixture (also called the Krichevski-Trofimov rule). Yet it differs importantly from the Jeffreys rule for small counts  $N_x$ .

However, when using two tilting parameters to adjust for relative importance of symbols within an alphabet, for example,  $Q_{a,b}$  in Section III-B, the predictive distribution does depend on  $b$ , i.e.,

$$\begin{aligned} & P(X_{i+1} = x | X_1, \dots, X_i) \\ &= \frac{e^{-\mathbf{1}_{\{x > L\}} b} (N_x^i + 1)^{N_x^i + 1} / N_x^{i N_x^i}}{\sum_{\tilde{x}=1}^m e^{-\mathbf{1}_{\{\tilde{x} > L\}} b} (N_{\tilde{x}}^i + 1)^{N_{\tilde{x}}^i + 1} / N_{\tilde{x}}^{i N_{\tilde{x}}^i}}. \end{aligned}$$

Hence, all symbols beyond  $L$  are discounted by an extra fact of  $e^{-b}$  when predicted by this rule.

## IV. APPLICATION

### A. Simulation

We first look at the performance of the tilted Stirling ratio distribution for algebraically decreasing counts with simulated data. The alphabet is partitioned into two subsets – the frequent symbols and the infrequent ones. The tilting parameter is chosen approximately according to the ratio of the number of symbols in a subset and their total count. The regret of assigning different number of symbols as ‘frequent’ ( $L$ ) is shown in Fig. 2.

We can see that more skewness pushes the optimizing  $L$  smaller.

### B. Real data

We also provide an example of using the tilted Stirling ratio distribution to code Chinese literature. The target book is an ancient collection of poems named 《诗经》, translated as the Classic of Poetry. It is the existing earliest collection of Chinese poetry and dates from the 10th to 7th centuries BC [18]. The book is downloaded freely from <http://wenku.baidu.com/>. Since many ancient words are rarely used today, the encoding is done in GB18030 [19], the largest Chinese coded character set. It contains 70244 characters, among which 2889 appear in the book with a total character count 39161. There are 792 characters appear once and 479 appear twice. The smallest regret happens at  $L = 2889$  which is the total number of characters appear.

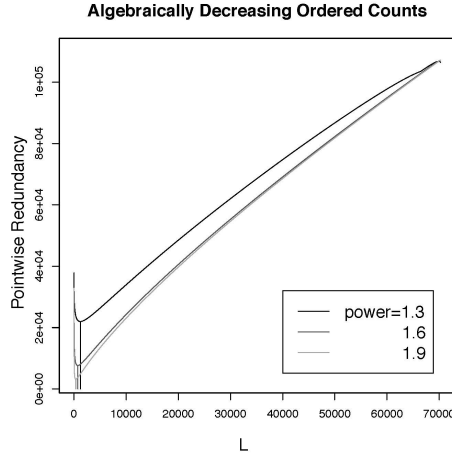


Fig. 2. Rregreterget of using tilted Stirling ratio distribution for algebraically decreasing counts.

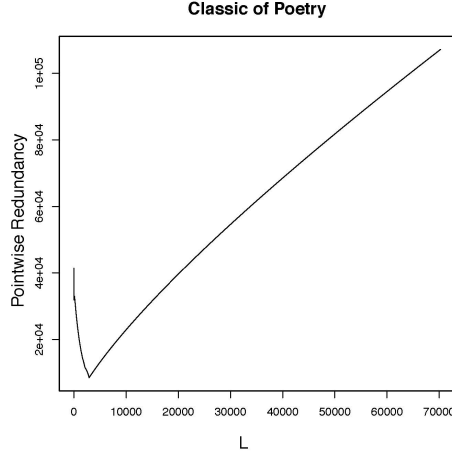


Fig. 3. Rregreterget of  $Q_{a,b}$  for  $L$  from 1 to  $m$ .

## V. DISCUSSION

We have introduced the use of independent tilted maximized Poisson likelihood distributions (also here called tilted Stirling ratio distributions)  $Q_a$  for coding the counts for independent random variables. The performance of the coding distribution is close to the minimax level. Actually, the difference between the regret and the minimax level is the probability assigned to the set with the observed total count by the tilted distribution with the optimal tilting parameter, i.e.

$$R(M_{cond}, \mathcal{P}_\Lambda^m, S_{m,n}) = R(Q_{a^*}, \mathcal{P}_\Lambda^m, S_{m,n}) + \log Q_{a^*}(S_{m,n}).$$

The optimal tilting parameter  $a^*$  minimizes the difference among all possible  $a$ . Since  $M_{cond}$  reproduces the Shtarkov NML distribution for the multinomial family of distributions on counts, it is the exact pointwise minimax strategy. As shown in this paper, our findings about the regret produced by the distribution  $Q_a$ , taken together with earlier work [2][3][12][8], show that the difference is no larger than about  $\log n$  in small alphabet cases, and about  $\frac{1}{2} \log n$  for moderate or large alphabets. The probability  $Q_a(S_{m,n})$  is the probability distribution for the total count  $N$  evaluated at  $N = n$  as induced by our distribution  $Q_a$ . Further analysis could be done to characterize this distribution of the total count more precisely.

## APPENDIX A

**Fact 1.** For any  $a > 0$ ,

$$\frac{1}{\sqrt{2\pi}} \int_0^1 t^{-\frac{1}{2}} e^{-at} dt < \sqrt{\frac{2}{\pi}}.$$

*Proof:*

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_0^1 t^{-\frac{1}{2}} e^{-at} dt &\stackrel{u=at}{=} \frac{1}{\sqrt{2\pi}} \int_0^a \left(\frac{u}{a}\right)^{-\frac{1}{2}} e^{-u} \frac{1}{a} du \\ &= \frac{1}{\sqrt{2\pi a}} \int_0^a u^{-\frac{1}{2}} e^{-u} du \end{aligned}$$

The integrand is smaller than  $u^{-\frac{1}{2}}$  on  $[0, a]$ , so the integral is upper bounded by

$$\frac{1}{\sqrt{2\pi a}} \int_0^a u^{-\frac{1}{2}} du = \sqrt{\frac{2}{\pi}}.$$

■

**Fact 2.** For any  $a > 0$ ,

$$\sum_{k=1}^{\infty} \frac{k^{-\frac{1}{2}}}{\sqrt{2\pi} e^{r_k}} e^{-ak} \geq \frac{1}{\sqrt{2\pi}} \int_1^{\infty} t^{-\frac{1}{2}} e^{-at} dt$$

when  $\frac{1}{12k+1} \leq r_k \leq \frac{1}{12k}$ .

*Proof:* It suffice to show

$$\sum_{k=1}^{\infty} \frac{k^{-\frac{1}{2}}}{e^{\frac{1}{12k}}} e^{-ak} \geq \int_1^{\infty} t^{-\frac{1}{2}} e^{-at} dt \quad (10)$$

As demonstrated in Fig. 4, the curve is bounded between the two step functions representing the two functions  $f(k) = k^{-1/2} e^{-ak}$  (the solid line), and  $g(k) = (k+1)^{-1/2} e^{-a(k+1)}$  (the dashed line). Note that  $t^{-\frac{1}{2}} e^{-at}$  is convex in  $t$ , hence the area above the curve is larger than the area below the curve in the rectangles between the two step functions. Although the (unnormalized) tilting probability is shrunk by an extra factor of  $e^{\frac{1}{12k}}$ , as long as it does not drag the step function down below the mid-point of the rectangle, inequality (10) still stands. It remains to show

$$\frac{e^{\frac{1}{12k}} - 1}{e^{\frac{1}{12k}}} \leq \frac{1}{2} \left( \frac{k^{-\frac{1}{2}} e^{-ak} - (k+1)^{-\frac{1}{2}} e^{-a(k+1)}}{k^{-\frac{1}{2}} e^{-ak}} \right),$$

where the left hand side is the part dragged down by the term  $e^{\frac{1}{12k}}$  as a portion of the solid line step function, and the right hand side is half of the rectangle between the two step functions as a portion of the same step function. Rearranging the terms and we actually have the following inequality holds for each  $k \geq 1$  and  $a > 0$ ,

$$\left( 1 + \left( \frac{k}{k+1} \right)^{\frac{1}{2}} e^{-a} \right) e^{\frac{1}{12k}} \leq 2.$$

Therefore Inequality (10) follows. ■

**Lemma 1** (Bounds for  $C_a$ ). For any  $a > 0$ , the following bounds hold for  $C_a$

$$\max(1, 1 - \sqrt{\frac{2}{\pi}} + \frac{1}{\sqrt{2a}}) < C_a < 1 + \frac{1}{\sqrt{2a}}, \quad (11)$$

and

$$1 + e^{-(a+1)} < C_a < 1 + e^{-(a+1)} + \frac{1}{2\sqrt{\pi}} \frac{e^{-2a}}{1 - e^{-a}}. \quad (12)$$

*Proof:* For the upper bound,

$$C_a = \sum_{k=0}^{\infty} \frac{k^k e^{-k}}{k!} e^{-ak} \stackrel{(a)}{=} 1 + \sum_{k=1}^{\infty} \frac{k^{-\frac{1}{2}}}{\sqrt{2\pi} e^{r_k}} e^{-ak} \quad (13)$$

Here (a) is by Robbins' refinement of Stirling's approximation where  $\frac{1}{12k+1} < r_k < \frac{1}{12k}$ .

The sum can be bounded by a gamma integral, so

$$\begin{aligned} C_a &\leq 1 + \frac{1}{\sqrt{2\pi}} \int_0^{\infty} t^{-\frac{1}{2}} e^{-ta} dt \\ &\leq 1 + \frac{1}{\sqrt{2\pi}} \frac{\Gamma(\frac{1}{2})}{a^{\frac{1}{2}}} \\ &= 1 + \frac{1}{\sqrt{2a}}. \end{aligned}$$

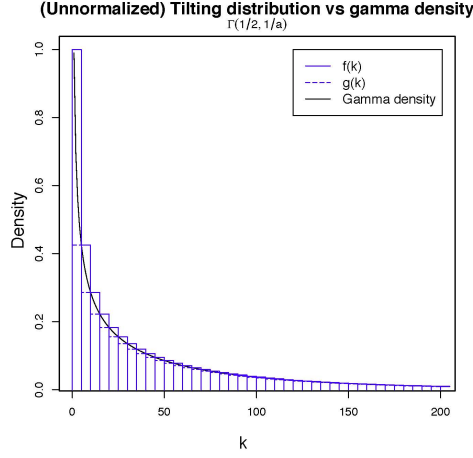


Fig. 4. tilted distribution vs  $\Gamma(\frac{1}{2}, \frac{1}{a})$  density.

Also, following expression (13),  $C_a$  has the following lower bound.

$$\begin{aligned}
 C_a &= 1 + \sum_{k=1}^{\infty} \frac{k^{-\frac{1}{2}}}{\sqrt{2\pi}e^{r_k}} e^{-ak} \\
 &\stackrel{(b)}{\geq} 1 - \sqrt{\frac{2}{\pi}} + \sqrt{\frac{2}{\pi}} + \frac{1}{\sqrt{2\pi}} \int_1^{\infty} t^{-\frac{1}{2}} e^{-at} dt \\
 &\stackrel{(c)}{>} 1 - \sqrt{\frac{2}{\pi}} + \frac{1}{\sqrt{2\pi}} \int_0^1 t^{-\frac{1}{2}} e^{-at} dt \\
 &\quad + \frac{1}{\sqrt{2\pi}} \int_1^{\infty} t^{-\frac{1}{2}} e^{-at} dt \\
 &= 1 - \sqrt{\frac{2}{\pi}} + \frac{1}{\sqrt{2\pi}} \int_0^{\infty} t^{-\frac{1}{2}} e^{-at} dt \\
 &= 1 - \sqrt{\frac{2}{\pi}} + \frac{1}{\sqrt{2a}}.
 \end{aligned}$$

Here again  $\frac{1}{12k+1} < r_k < \frac{1}{12k}$ , and inequality (b) is due to Fact 2 and inequality c is by Fact 1.

Note that inequality (11) is good for small  $a$ . For a moderately large ( $a > 0.2$ ), the following upper bound is better.

$$\begin{aligned}
 C_a &\leq 1 + e^{-(a+1)} + \sum_{k=2}^{\infty} \frac{1}{\sqrt{2\pi k}} e^{-ka} \\
 &< 1 + e^{-(a+1)} + \frac{1}{2\sqrt{\pi}} \frac{e^{-2a}}{1 - e^{-a}}.
 \end{aligned}$$

■

**Lemma 2.** For any  $a > 0$ ,

$$e^{-(a+1)} \leq \mathbf{E}_{P_a} N_1 \leq \frac{1}{2a}.$$

*Proof:* Let  $k^* = \min_{k \in \mathbf{N}_+} |k - \frac{1}{2a}|$ . We prove the upper bound by consider  $a$  within two different intervals. First, if

$a < e(\sqrt{\pi} - \sqrt{2})^2$ , we know

$$\begin{aligned}
& \sum_{k=1}^{\infty} \frac{k^{k+1} e^{-k}}{k!} e^{-ak} \\
&= \sum_{k=1}^{k^*-1} \frac{k^{k+1} e^{-k}}{k!} e^{-ak} + \sum_{k=k^*+1}^{\infty} \frac{k^{k+1} e^{-k}}{k!} e^{-ak} \\
&\quad + \frac{k^{*k^*+1} e^{-k^*}}{k^*!} e^{-ak^*} \\
&\stackrel{(a)}{\leq} \sum_{k=1}^{k^*-1} \frac{k^{1/2} e^{-ak}}{\sqrt{2\pi}} + \sum_{k=k^*+1}^{\infty} \frac{k^{1/2} e^{-ak}}{\sqrt{2\pi}} \\
&\quad + \frac{k^{*1/2} e^{-ak^*}}{\sqrt{2\pi}}
\end{aligned} \tag{14}$$

where (a) is an upper bound by Stirling's approximation.

Both sums in the last expression can be upper bounded by a gamma integral, and  $k^{*1/2} e^{-ak^*}$  is no larger than the maximum of the unnormalized  $\text{gamma}(3/2, 1/a)$  density, which is achieved at  $1/(2a)$ . Hence, we have the following upper bound for expression (14).

$$\begin{aligned}
& \int_0^{k^*} \frac{t^{1/2} e^{-at}}{\sqrt{2\pi}} dt + \int_{k^*}^{\infty} \frac{t^{1/2} e^{-at}}{\sqrt{2\pi}} dt + \frac{(1/2a)^{1/2} e^{-1/2}}{\sqrt{2\pi}} \\
&= \frac{\Gamma(3/2)}{a^{3/2} \sqrt{2\pi}} + \frac{(1/2a)^{1/2}}{\sqrt{2\pi} e} \\
&= \frac{1}{(2a)^{3/2}} + \frac{1}{\sqrt{2\pi} e} \frac{1}{(2a)^{1/2}}
\end{aligned}$$

Using this upper bound for  $C_a$ , we could prove an upper bound for the expected value.

$$\begin{aligned}
\mathbf{E}_{P_a} N_1 &= \sum_{k=1}^{\infty} \frac{k^{k+1} e^{-k}}{k! C_a} e^{-ak} \\
&\stackrel{(b)}{\leq} \frac{\frac{1}{(2a)^{3/2}} + \frac{1}{\sqrt{2\pi} e} \frac{1}{(2a)^{1/2}}}{\frac{1}{(2a)^{1/2}} + 1 - \sqrt{\frac{2}{\pi}}} \\
&= \frac{1}{2a} \underbrace{\left( \frac{\frac{1}{(2a)^{1/2}} + \frac{1}{\sqrt{2\pi} e} (2a)^{1/2}}{\frac{1}{(2a)^{1/2}} + 1 - \sqrt{\frac{2}{\pi}}} \right)}_{(A)}
\end{aligned}$$

The lower bound for the denominator in (b) is attributed to Lemma 1. A little algebra can show that term (A) is no larger than 1 when  $a$  is restricted to  $(0, e(\sqrt{\pi} - \sqrt{2})^2)$ .

If  $a > e(\sqrt{\pi} - \sqrt{2})^2$ , we have  $k^* = 1$ .

$$\begin{aligned}
& \sum_{k=1}^{\infty} \frac{k^{k+1} e^{-k}}{k!} e^{-ak} \\
&\leq \sum_{k=1}^{\infty} \frac{k^{1/2} e^{-ak}}{\sqrt{2\pi}} \\
&\stackrel{(c)}{\leq} \frac{1}{\sqrt{2\pi}} \left( \int_0^{\infty} t^{1/2} e^{-at} dt + \frac{1}{2} e^{-a} \right) \\
&= \frac{1}{\sqrt{2\pi}} \left( \frac{\Gamma(3/2)}{a^{3/2}} + \frac{1}{2} e^{-a} \right) \\
&= \frac{1}{(2a)^{3/2}} + \frac{1}{2\sqrt{2\pi}} e^{-a}
\end{aligned}$$

where (c) is because the difference between  $\int_{t=0}^1 t^{1/2} e^{-at} dt$  and  $e^{-a}$  is less than  $\frac{1}{2} e^{-a}$ .

By this upper bound for the numerator and Lemma 1 again,

$$\begin{aligned} \mathbf{E}_{P_a} N_1 &\leq \frac{\frac{1}{(2a)^{3/2}} + \frac{1}{2\sqrt{2\pi}} e^{-a}}{\frac{1}{(2a)^{1/2}} + 1 - \sqrt{\frac{2}{\pi}}} \\ &= \frac{1}{2a} \underbrace{\left( \frac{\frac{1}{(2a)^{1/2}} + \frac{1}{\sqrt{2\pi}} a e^{-a}}{\frac{1}{(2a)^{1/2}} + 1 - \sqrt{\frac{2}{\pi}}} \right)}_{(B)}. \end{aligned}$$

Term (B) is no larger than 1 because  $\frac{1}{\sqrt{2\pi}} a e^{-a} \leq 1 - \sqrt{\frac{2}{\pi}}$  for all  $a$ .

For the lower bound,

$$\begin{aligned} \mathbf{E}_{p_a} N_1 &= \sum_{k=1}^{\infty} \frac{k^{k+1} e^{-k}}{k! C_a} e^{-ak} \\ &= \frac{e^{-(a+1)} \left( \sum_{k=1}^{\infty} \frac{k^k e^{-(k-1)}}{(k-1)!} e^{-a(k-1)} \right)}{C_a} \\ &\stackrel{l=k-1}{=} \frac{e^{-(a+1)} \left( \sum_{l=0}^{\infty} \frac{(l+1)^{l+1} e^{-l}}{l!} e^{-al} \right)}{C_a} \\ &= e^{-(a+1)} \underbrace{\left( \frac{\sum_{l=0}^{\infty} \frac{(l+1)^{l+1} e^{-l}}{l!} e^{-al}}{\sum_{k=0}^{\infty} \frac{k^k e^{-k}}{k!} e^{-ak}} \right)}_{(C)} \\ &\stackrel{(d)}{\geq} e^{-(a+1)} \end{aligned} \tag{15}$$

Here inequality (d) is because term (C) is above 1. Hence the upper bound is deduced. ■

## APPENDIX B PROOF OF THEOREM 1

*Proof:* It remains to show the two lower bounds in expression (5) and (6). In both cases we need a lower bound for  $na^* \log e + m \log C_{a^*}$ , and we do it by lower bounding  $a^*$  and  $C_{a^*}$ , respectively. Let  $\tilde{a} = \frac{m}{2n}$ .

- Bounds for  $a^*$

We know  $a^*$  is the solution for the following equation.

$$\mathbf{E}_{P_{a^*}} N_1 = \frac{n}{m}$$

By Lemma 2, we have

$$\frac{1}{2a^*} \geq \frac{n}{m}$$

That gives

$$a^* \leq \frac{m}{2n} = \tilde{a} \tag{16}$$

Since  $C_a$  is decreasing in  $a$ , we have

$$C_{a^*} \geq C_{\tilde{a}} > \frac{1}{\sqrt{2\tilde{a}}} = \sqrt{\frac{n}{m}}.$$

For any  $j \in \{1, \dots, m\}$ , and  $a > 0$ , we have

$$\begin{aligned} \mathbf{E}_{P_a} N_1 &= \sum_{k=1}^{\infty} \frac{k^{k+1} e^{-k}}{k! C_a} e^{-ak} \\ &\stackrel{(a)}{\geq} \frac{\sum_{k=1}^{\infty} \frac{k^{k+1} e^{-k}}{k!} e^{-ak}}{1 + \frac{1}{\sqrt{2a}}} \\ &\stackrel{(b)}{=} \frac{\sum_{k=1}^{\infty} \frac{k^{\frac{1}{2}}}{\sqrt{2\pi} e^{r_k}} e^{-ak}}{1 + \frac{1}{\sqrt{2a}}} \end{aligned} \tag{17}$$

Here (a) is attributed to inequality (11), step (b) is by Stirling's approximation, and  $\frac{1}{12k+1} < r_k < \frac{1}{12k}$ . Pick  $k_1 = a^{-1/3}$ , then the numerator of expression (17) can be lower bounded by

$$\begin{aligned} & \sum_{k=\lfloor k_1 \rfloor}^{\infty} \frac{k^{1/2}}{\sqrt{2\pi} e^{r_k}} e^{-ak} \\ & \geq \sum_{k=\lfloor k_1 \rfloor}^{\infty} \frac{k^{1/2}}{\sqrt{2\pi} e^{\frac{1}{12\lfloor k_1 \rfloor}}} e^{-ak} \\ & \geq \frac{1}{\sqrt{2\pi} e^{\frac{1}{12(k_1-1)}}} \int_{\lfloor k_1 \rfloor}^{\infty} t^{1/2} e^{-at} dt \end{aligned}$$

Taking the integral from 0 to  $\infty$  and subtracting the part from 0 to  $k_1$  yields the lower bound

$$\begin{aligned} & \frac{1}{\sqrt{2\pi} e^{\frac{1}{12(k_1-1)}}} \left( \frac{\Gamma(3/2)}{a^{3/2}} - \int_0^{k_1} t^{1/2} e^{-at} dt \right) \\ & \geq \frac{1}{\sqrt{2\pi} e^{\frac{1}{12(k_1-1)}}} \left( \frac{\Gamma(3/2)}{a^{3/2}} - \int_0^{k_1} t^{1/2} dt \right) \\ & = \frac{1}{\sqrt{2\pi} e^{\frac{1}{12(k_1-1)}}} \left( \frac{\Gamma(3/2)}{a^{3/2}} - \frac{2}{3a^{1/2}} \right). \end{aligned}$$

Write  $r_a = \frac{1}{12(k_1-1)} = \frac{a^{1/3}}{12(1-a^{1/3})}$ . By the above calculation, we have a lower bound for the expectation under the tilting distribution. For  $a^*$ ,

$$\frac{\frac{1}{\sqrt{2\pi} e^{r_{a^*}}} \left( \frac{\Gamma(3/2)}{a^{*3/2}} - \frac{2}{3a^{*1/2}} \right)}{1 + \frac{1}{\sqrt{2a^*}}} \leq \mathbf{E}_{a^*} N_1 = \frac{n}{m}.$$

Arranging the terms, we have

$$\begin{aligned} \frac{1}{2a^*} & \leq \frac{n}{m} \left( 1 + \sqrt{2a^*} \right) e^{r_{a^*}} + \frac{2}{3\sqrt{\pi}} \\ & \stackrel{(c)}{\leq} \frac{n}{m} \left( 1 + \sqrt{2\tilde{a}} \right) e^{r_{\tilde{a}}} + \frac{2}{3\sqrt{\pi}} \end{aligned}$$

Here (c) is because  $a^* \leq \tilde{a}$  by inequality (16). So,

$$a^* \geq \frac{\tilde{a}}{\left( 1 + \sqrt{2\tilde{a}} \right) e^{r_{\tilde{a}}} + \frac{4}{3\sqrt{\pi}} \tilde{a}}$$

By Taylor expansion, this is no smaller than

$$\begin{aligned} & \frac{\tilde{a}}{\left( 1 + \sqrt{2\tilde{a}} \right) (1 + r_{\tilde{a}} + O(r_{\tilde{a}}^2)) + \frac{4}{3\sqrt{\pi}} \tilde{a}} \\ & = \tilde{a} \left( 1 - \frac{r_{\tilde{a}} + \sqrt{2\tilde{a}} + \sqrt{2\tilde{a}} r_{\tilde{a}} + \frac{4}{3\sqrt{\pi}} \tilde{a} + O(r_{\tilde{a}}^2)}{\left( 1 + \sqrt{2\tilde{a}} \right) (1 + r_{\tilde{a}} + O(r_{\tilde{a}}^2)) + \frac{4}{3\sqrt{\pi}} \tilde{a}} \right) \\ & \geq \tilde{a} \left( 1 - r_{\tilde{a}} - \sqrt{2\tilde{a}} - \sqrt{2\tilde{a}} r_{\tilde{a}} - \frac{4}{3\sqrt{\pi}} \tilde{a} - O(r_{\tilde{a}}^2) \right) \end{aligned}$$

When  $m = o(n)$ ,  $r_{\tilde{a}}$  is the leading term, so

$$a^* \geq \tilde{a} (1 - O(r_{\tilde{a}})) = \frac{m}{2n} \left( 1 - O\left( \left( \frac{m}{n} \right)^{\frac{1}{3}} \right) \right)$$

As a result,

$$na^* \log e \geq \left( 1 - O\left( \left( \frac{m}{n} \right)^{\frac{1}{3}} \right) \right) \frac{m}{2} \log e$$

Hence we get inequality (5).

The above lower bound works when  $a^*$  is small (i.e., when  $m$  is small compared to  $n$ ), yet when it is large, the following bound is better. Let  $a_0 = \ln \frac{m}{ne}$ .



For large  $m$ ,  $a_0$  is the solution of the following equation.

$$\mathbf{E}_{P_a} N_1 = \frac{n}{m}$$

That is

$$\sum_{k=1}^{\infty} \frac{k^{k+1} e^{-k}}{k! C_{a_0}} e^{-a_0 k} = \frac{n}{m}$$

The left hand side is lower bounded by  $e^{-(a_0+1)}$  by Lemma 2. Hence we have

$$\begin{aligned} e^{-(a^*+1)} &\leq \frac{n}{m} \\ e^{a^*} &\geq \frac{m}{ne} = e^{a_0} \\ a^* &\geq a_0 \end{aligned}$$

Thus,

$$na^* \log e \geq na_0 \log e = n \log \frac{m}{ne}$$

- Bounds for  $C_{a^*}$

Now we want to lower bound  $C_{a^*}$ . Recall inequality (15), let term  $(C)$  be defined as

$$r_a = \frac{\sum_{l=0}^{\infty} (l+1)^{l+1} e^{-l} e^{-al} / l!}{\sum_{k=0}^{\infty} k^k e^{-k} e^{-ak} / k!}.$$

We have

$$r_{a^*} e^{-(a^*+1)} = \mathbf{E}_{P_{a^*}} N_j = \frac{n}{m} = e^{-(a_0+1)}.$$

It gives

$$e^{-(a^*+1)} = \frac{e^{-(a_0+1)}}{r_{a^*}}.$$

By definition,

$$C_{a^*} \geq 1 + e^{-(a^*+1)} = 1 + \frac{e^{-(a_0+1)}}{r_{a^*}}.$$

Hence, we have

$$\begin{aligned} \frac{1}{r_{a^*}} &\geq \frac{1 + e^{-(a^*+1)}}{1 + 4e^{-(a^*+1)} + \frac{27}{2} \frac{e^{-2(a^*+1)}}{1 - e^{-(a^*+1)}}} \\ &= 1 - \frac{3e^{-(a^*+1)} + \frac{27}{2} \frac{e^{-2(a^*+1)}}{1 - e^{-(a^*+1)}}}{1 + 4e^{-(a^*+1)} + \frac{27}{2} \frac{e^{-2(a^*+1)}}{1 - e^{-(a^*+1)}}} \\ &\geq 1 - 3e^{-(a^*+1)} - \frac{27}{2} \frac{e^{-2(a^*+1)}}{1 - e^{-(a^*+1)}} \\ &= 1 - O\left(e^{-(a^*+1)}\right) \\ &\geq 1 - O\left(e^{-(a_0+1)}\right). \end{aligned}$$

So

$$C_{a^*} \geq 1 + \left(1 - O\left(e^{-(a_0+1)}\right)\right) e^{-(a_0+1)}$$

And we derive a lower bound

$$m \log C_{a^*} \geq m \log \left(1 + \left(1 - O\left(\frac{n}{m}\right)\right) \frac{n}{m}\right).$$

Therefore, inequality (6) follows. ■

## APPENDIX C REDUNDANCY

**Theorem 0.** Let  $M$  denote the Stirling ratio distribution as defined before, and  $M_{cond}$  be the measure  $M$  conditional on the observed value  $\frac{1}{m} \sum_{j=1}^m h(N_j) = \alpha$ , where  $h$  is a function of the data, and let  $P_a$  be the tilted distribution with parameter  $a$ , chosen by the condition  $\mathbf{E}_{P_a} h(N_1) = \alpha$ , and  $C_\alpha$  be a class of distributions with the expected value equal to the observed

$$C_\alpha = \{P : \mathbf{E}_P h(N_1) = \alpha\}.$$

Similar to what has been shown in [4], [5], and [6] for i.i.d random variables,  $Q_a = \otimes_{j=1}^m P_a$  is the information projection of  $M$  on  $C_\alpha$ . In fact,

$$\begin{aligned} D(M_{cond}||M) &= D(M_{cond}||Q_a) + D(Q_a||M) \\ &\geq D(Q_a||M). \end{aligned}$$

Theorem 0 says the tilted distribution is closest to the original distribution in relative entropy among all distributions with the expected value of a function equal to the observed value. Hence it is the redundancy minimizing distribution over the class of distributions with a given moment condition. In particular, if  $h(x) = x$ , the condition is made on the total count.

**Theorem 4.** Let  $X_1, \dots, X_N$  be generated from an alphabet  $\mathcal{A}$  of size  $m$ , where  $N$  is also random. Let  $\lambda_{sum}$  denote the mean of  $N$ , i.e.  $\lambda_{sum} := \sum_{j=1}^m \lambda_j$ , with  $\lambda_{sum}$  known and  $\lambda_j > 0$  for all  $j$ . Let  $\mathcal{P}_{\lambda_{sum}}^m$  denote the class of distributions on  $N_1, \dots, N_m$  with expected total count equal to  $\lambda_{sum}$ . The redundancy by using a tilted distribution  $Q_a$  is mainly

$$(A) = \left( -\frac{m}{2} + a\lambda_{sum} \right) \log e + m \log C_a,$$

with the error bounded by

$$\sum_{j=1}^m \left( \frac{1}{3\lambda_j^2} + \frac{5}{6\lambda_j} \right) \log e.$$

The error is small if no  $\lambda_j$  is too small. Moreover, the minimizer of the redundancy is  $a^*$ . Therefore  $P_{a^*}$  is an minimax strategy within the class of distributions with mean  $\lambda_{sum}$ . The minimax redundancy has the following approximations according to the magnitude of  $m$  and  $\lambda_{sum}$ .

When  $m = o(\lambda_{sum})$ , term (A) satisfies the following inequality

$$0 \leq \left| (A) - \frac{m}{2} \log \frac{\lambda_{sum}}{m} \right| \leq m \log \left( 1 + \sqrt{\frac{m}{\lambda_{sum}}} \right). \quad (18)$$

When  $\lambda_{sum} = o(m)$ , term (A) satisfies the following inequality

$$\begin{aligned} & m \log \left( 1 + \frac{\lambda_{sum}}{m} \right) - \lambda_{sum} \log e \\ & \leq \left| (A) - \left( \lambda_{sum} \log \frac{m}{\lambda_{sum}} - \frac{m}{2} \log e \right) \right| \\ & \leq \frac{1}{2\sqrt{\pi}} \frac{\lambda_{sum}^2 e^2}{m - \lambda_{sum} e} \log e. \end{aligned} \quad (19)$$

*Proof:* The first part of the proof follows Lemma 3 in [3], and the second part resembles proof of Theorem 1.

$$\begin{aligned} & \mathbf{E}_\lambda \ln \frac{\prod_{j=1}^m P_{\lambda_j}(N_j)}{Q_a(N)} \\ &= \sum_{j=1}^m (\lambda_j \ln \lambda_j) - \sum_{j=1}^m \mathbf{E}_{\lambda_j} (N_j \ln N_j) + a \sum_{j=1}^m \lambda_j \\ & \quad + m \ln C_a \end{aligned}$$

Following Lemma 3 in [3], by Taylor's expansion, for each  $j$ ,

$$\begin{aligned} & \mathbf{E}_{\lambda_j} (N_j \ln N_j) \\ & \geq \lambda_j \ln \lambda_j + \mathbf{E}_{\lambda_j} (N_j - \lambda_j)(1 + \ln \lambda_j) \\ & \quad + \mathbf{E}_{\lambda_j} \frac{1}{2} (N_j - \lambda_j)^2 \frac{1}{\lambda_j} + \frac{1}{6} \mathbf{E}_{\lambda_j} (N_j - \lambda_j)^3 \left( -\frac{1}{\lambda_j^2} \right) \\ & = \lambda_j \ln \lambda_j + \frac{1}{2} - \frac{1}{6\lambda_j}. \end{aligned}$$

We also know by Jensen's Inequality that

$$\mathbf{E}_{\lambda_j}(N_j \ln N_j) \geq \lambda_j \ln \lambda_j.$$

Hence,

$$\mathbf{E}_{\lambda_j}(N_j \ln N_j) \geq \lambda_j \ln \lambda_j + \frac{1}{2} + \max\left(-\frac{1}{6\lambda_j}, -\frac{1}{2}\right).$$

And

$$\begin{aligned} & \mathbf{E}_{\lambda_j}(N_j \ln N_j) \\ & \leq \lambda_j \ln \lambda_j + (\mathbf{E}_{\lambda_j} N_j - \lambda_j)(1 + \ln \lambda_j) \\ & \quad + \frac{\mathbf{E}_{\lambda_j}(N_j - \lambda_j)^2}{2\lambda_j} - \frac{\mathbf{E}_{\lambda_j}(N_j - \lambda_j)^3}{6\lambda_j^2} \\ & \quad + \frac{\mathbf{E}_{\lambda_j}(N_j - \lambda_j)^4}{3\lambda_j^3} \\ & = \lambda_j \ln \lambda_j + \frac{1}{2} + \frac{1}{3\lambda_j^2} + \frac{5}{6\lambda_j}. \end{aligned}$$

Therefore,

$$\begin{aligned} & - \left( \sum_{j=1}^m \frac{1}{3\lambda_j^2} + \frac{5}{6\lambda_j} \right) \\ & \leq \mathbf{E}_{\underline{\lambda}} \ln \frac{\prod_{j=1}^m P_{\lambda_j}(N_j)}{Q_a(\underline{N})} \\ & \quad - \left( -\frac{m}{2} + a \sum_{j=1}^m \lambda_j + m \ln C_a \right) \\ & \leq \min \left( \sum_{j=1}^m \frac{1}{6\lambda_j}, \frac{m}{2} \right). \end{aligned}$$

The fact that  $a^*$  is the minimizer can be easily seen by taking partial derivative with respect to  $a$  for term (A). The two inequalities are attributed to Lemma 1, by picking  $a = m/(2\lambda_{sum})$  and  $a = \ln(m/\lambda_{sum}e)$  respectively. ■

#### APPENDIX D PROOF OF THEOREM 3

*Proof:* The MLE for an envelope class is the following

$$\hat{\lambda}_j = \arg \sup_{\lambda_j \leq nf(j)} P_{\lambda_j}(N_j) = N_j \wedge nf(j).$$

We formulate a tilted distribution by multiplying the exponential tilting factor  $e^{-aN_j}$  for each  $j \in \{1, \dots, m\}$  and normalize it.

$$P_a(N_j) = \begin{cases} \frac{N_j^{N_j} e^{-N_j}}{N_j!} \frac{e^{-aN_j}}{C_{a,j}} & \text{if } N_j \leq nf(j) \\ \frac{(nf(j))^{N_j} e^{-nf(j)}}{N_j!} \frac{e^{-aN_j}}{C_{a,j}} & \text{if } N_j > nf(j) \end{cases}$$

where  $C_{a,j} = \sum_{N_j \leq nf(j)} \frac{N_j^{N_j} e^{-N_j}}{N_j!} e^{-aN_j} + \sum_{N_j > nf(j)} \frac{(nf(j))^{N_j} e^{-nf(j)}}{N_j!} e^{-aN_j}$ .

The regret of using independent  $P_a$  for each  $N_j$  in  $\underline{N} \in S_{m,n}$  is

$$\log \prod_{j=1}^m \frac{P_{\hat{\lambda}_j}(N_j)}{P_a(N_j)} = na \log e + \sum_{j=1}^m \log C_{a,j}. \quad (20)$$

Again,  $a^*$  minimizes expression (20).

For each  $j$  and any positive  $a$ ,

$$\begin{aligned} C_{a,j} &= \sum_{N_j \leq \lfloor nf(j) \rfloor} \frac{N_j^{N_j} e^{-N_j}}{N_j!} e^{-aN_j} \\ & \quad + \sum_{N_j > nf(j)} \frac{(nf(j))^{N_j} e^{-nf(j)}}{N_j!} e^{-aN_j}. \end{aligned}$$

The sum only depends on the envelope function  $f(j)$  for given  $a$  and  $j$ .

Since  $(nf(j))^x e^{-nf(j)} \leq x^x e^{-x}$  for all  $x > 0$ , for any symbol  $j$  with  $N_j > nf(j)$ , we have

$$\frac{(nf(j))^{N_j} e^{-nf(j)}}{N_j!} e^{-aN_j} \leq \frac{N_j^{N_j} e^{-N_j}}{N_j!} e^{-aN_j}.$$

Hence we have,

$$\begin{aligned} C_{a,j} &\leq \sum_{N_j=0}^{\infty} \frac{N_j^{N_j} e^{-N_j}}{N_j!} e^{-aN_j} \\ &\leq 1 + \sum_{N_j=1}^{\infty} \frac{N_j^{N_j} e^{-N_j}}{N_j!} e^{-aN_j} \\ &\stackrel{(a)}{\leq} 1 + \sum_{N_j=1}^{\infty} \frac{1}{\sqrt{2\pi}} N_j^{-1/2} e^{-aN_j} \\ &\leq 1 + \frac{1}{\sqrt{2\pi}} \int_0^{\infty} t^{-\frac{1}{2}} e^{-at} dt \\ &= 1 + \sqrt{\frac{1}{2a}} \end{aligned}$$

where (a) is by Stirling's approximation. This is a good bound when  $nf(j)$  is big.

However, if  $nf(j)$  is small, the following upper bound is better. For  $N_j \leq \lfloor nf(j) \rfloor$ ,

$$\begin{aligned} \sum_{N_j \leq \lfloor nf(j) \rfloor} \frac{N_j^{N_j} e^{-N_j}}{N_j!} e^{-aN_j} &\leq \sum_{N_j \leq \lfloor nf(j) \rfloor} \frac{N_j^{N_j}}{N_j!} \\ &\leq \sum_{N_j \leq \lfloor nf(j) \rfloor} \frac{(nf(j))^{N_j}}{N_j!}. \end{aligned}$$

For the second partial sum, we also have

$$\begin{aligned} &\sum_{N_j > nf(j)} \frac{(nf(j))^{N_j} e^{-nf(j)}}{N_j!} e^{-aN_j} \\ &\leq \sum_{N_j > nf(j)} \frac{(nf(j))^{N_j}}{N_j!}. \end{aligned}$$

Deduce,

$$C_{a,j} \leq \sum_{N_j=0}^{\infty} \frac{(nf(j))^{N_j}}{N_j!} = e^{nf(j)}.$$

Hence for any given  $a, j$  and  $L \in \{1, 2, \dots, m\}$ , the following upper bound holds.

$$\begin{aligned} &na \log e + \sum_{j=1}^m \log C_{a,j} \\ &\leq na \log e + \log \left( \prod_{j=1}^L \left( 1 + \sqrt{\frac{1}{2a}} \right) \prod_{j=L+1}^m \left( e^{nf(j)} \right) \right) \\ &= na \log e + L \log \left( 1 + \sqrt{\frac{1}{2a}} \right) \\ &\quad + \left( \sum_{j=L+1}^m nf(j) \right) \log e. \end{aligned}$$

Let  $a = \frac{L}{2(n - \sum_{j>L} nf(j))}$ , the result follows. ■

## ACKNOWLEDGMENT

The authors would like to thank Prof. Mokshay Madiman and Prof. Joseph Chang for their helpful advice and consistent support. They are also grateful to the family-like department and classmates, which makes all of these possible.

## REFERENCES

- [1] (2014, Jan). [Online]. Available: [http://en.wikipedia.org/wiki/Chinese\\_characters](http://en.wikipedia.org/wiki/Chinese_characters)
- [2] Y. M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, no. 3, pp. 175–186, 1987.
- [3] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Transactions on Information Theory*, vol. 43, pp. 646–657, May 1997.
- [4] I. Csiszar, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, Feb 1975.
- [5] —, "Sanov property, generalized I-projection and a conditional limit theorem," *The Annals of Probability*, vol. 12, no. 3, pp. 768–793, Jan 1984.
- [6] J. V. Campenhout and T. Cover, "Maximum entropy and conditional probability," *IEEE Transactions on Information Theory*, vol. 27, no. 4, July 1981.
- [7] A. Orlistsky, N. P. Santhanam, and J. Zhang, "Always good turing: Asymptotically optimal probability estimation," *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, 2003.
- [8] W. Szpankowski and M. J. Weinberger, "Minimax redundancy for large alphabets," *Information Theory Proceedings*, June 2010.
- [9] L. A. Adamic, Zipf, power-laws, and pareto - a ranking tutorial. [Online]. Available: <http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html>
- [10] Y. M. Shtarkov, T. J. Tjalkens, and F. M. J. Willems, "Multi-alphabet universal coding of memoryless sources," *Problems of Information Transmissions*, vol. 31, no. 2, pp. 114–127, 1995.
- [11] Q. Xie and A. R. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 431–445, March 2000.
- [12] A. Orlistsky and N. P. Santhanam, "Speaking of infinity," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2215–2230, October 2004.
- [13] A. G. S. Boucheron and E. Gassiat, "Coding on countably infinite alphabets," *IEEE Transactions on Information Theory*, vol. 55, no. 1, Jan 2009.
- [14] D. Bontemps, "Universal coding on infinite alphabets: exponentially decreasing envelopes," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1466–1478, 2011.
- [15] S. Kullback, *Information Theory and Statistics*. Wiley, New York, 1959.
- [16] P. Bartlett, P. Grunwald, P. Harremoës, F. Hedayati, and W. Kotlowski, "Horizon-independent optimal prediction with log-loss in exponential families," *arXiv preprint arXiv:1305.4324*, 2013.
- [17] T. Roos and J. Rissanen, "On sequentially normalized maximum likelihood models," in *Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08)*, August 2008.
- [18] (2013, September). [Online]. Available: [https://en.wikipedia.org/wiki/Classic\\_of\\_Poetry](https://en.wikipedia.org/wiki/Classic_of_Poetry)
- [19] (2013, September). [Online]. Available: [http://zh.wikipedia.org/wiki/GB\\_18030](http://zh.wikipedia.org/wiki/GB_18030)